Mozer, M. C. (1988). A connectionist model of selective attention in visual perception. In V. L. Patel & G. J. Groen (Eds.), Proceedings of the Tenth Annual Conference of the Cognitive Science Society (pp. 195-201). Hillsdale, NJ: Erlbaum Associates.

A connectionist model of selective attention in visual perception

Michael C. Mozer

Institute of Cognitive Science
University of Colorado, Boulder

This paper describes a model of selective attention that is part of a connectionist object recognition system called MORSEL. MORSEL is capable of identifying multiple objects presented simultaneously on its "retina," but because of capacity limitations, MORSEL requires attention to prevent it from trying to do too much at once. Attentional selection is performed by a network of simple computing units that constructs a variable-diameter "spotlight" on the retina, allowing sensory information within the spotlight to be preferentially processed. Simulations of the model demonstrate that attention is more critical for less familiar items and that attention can be used to reduce inter-item crosstalk. The model suggests four distinct roles of attention in visual information processing, as well as a novel view of attentional selection that has characteristics of both early and late selection theories.

Few would argue that the visual system is unlimited in its capacity for processing sensory information. Some means of selective and sequential analysis is required. This is the primary function of attention: to control the amount and the temporal order of information flowing through the visual system. Any complete model of visual information processing must thus address the issue of attention. In this paper, I describe an attentional mechanism designed for a connectionist model of two-dimensional object recognition called *MORSEL* (Mozer, 1987a, b). MORSEL is capable of identifying multiple objects presented simultaneously on its "retina," but because of capacity limitations, MORSEL requires an attentional mechanism to prevent it from trying to do too much at once and making errors.

Briefly, MORSEL (Figure 1) consists of four components: (1) a set of processing modules that analyze objects along various attribute dimensions; (2) a network that constructs a consistent interpretation of the perceptual data provided by these modules (the *pull-out net*); (3) an *attentional mechanism* (*AM* for short) that guides the efforts of the modules; and (4) a *visual short-term memory* that holds object descriptions. To illustrate the typical operation of the system, consider a simple example in which MORSEL is shown a display containing two colored letters, a red X and a blue T. These letters will cause a pattern of activity on MORSEL's retina, which serves as input to each of the processing modules as well as to the AM. The AM then focuses on one retinal region, say the location of the red X. Information from that region is processed by each module. One module extracts shape information, identifying the object as an "x" or possibly a "y," another extracts color information, identifying the object as being red. The pull-out net then selects the most plausible interpretation of each module's output, in this case "x" and "red." The representation at this level of the system encodes attributes of the visual object *without regard to location.* Location information is recovered from the AM, which indicates the current location of focus. Shape, color, and location information are then bound together and stored in the short-term memory. Next, attention shifts to the blue T, and this process repeats.

I have built a computer simulation of MORSEL with one module elaborated in detail — a letter and word recognition system called *BLIRNET*. BLIRNET has been trained to recognize letters and words in arbitrary retinal locations, and is able to recognize several items simultaneously, although interactions within the network limit the number of items that can be accurately processed. BLIRNET is a hierarchical multi-layered network. Its input layer is a retinotopic feature map arranged in a 36×6 spatial array, with detectors for five feature types at each point in the array (line segments at four orientations and line-segment terminator detectors). Letters of the alphabet are encoded as an activity pattern over a 3×3 retinal region. BLIRNET's output layer contains *letter-cluster detectors*, which respond to single

My thanks to Don Norman, Hal Pashler, and Geoff Hinton for their guidance, and to Steve Nowlan for helpful comments on an earlier draft. This work was supported by grant 87-2-36 from the Alfred P. Sloan Foundation to Geoffrey Hinton, Contract N00014-85-C-0133 NR 667-541 with the Personnel and Training Research Programs of the Office of Naval Research, and a grant from the System Development Foundation to Donald Norman and David Rumelhart.

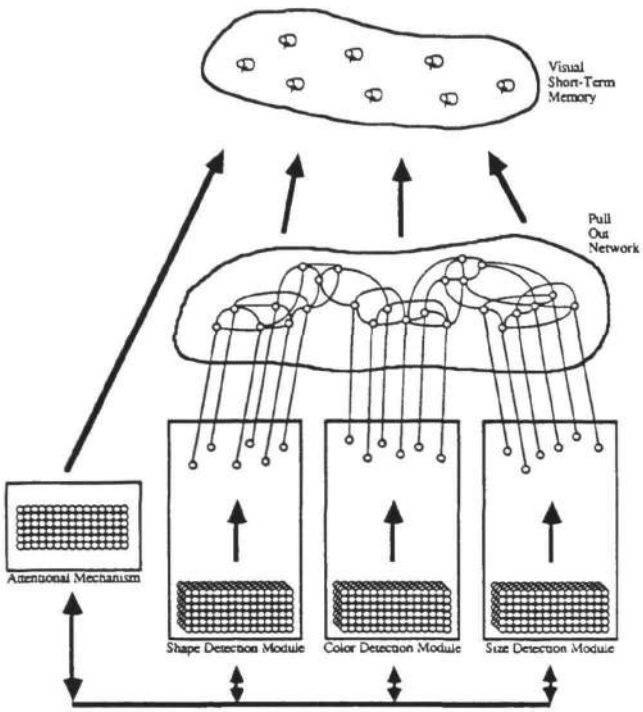

Figure 1. A sketch of MORSEL.

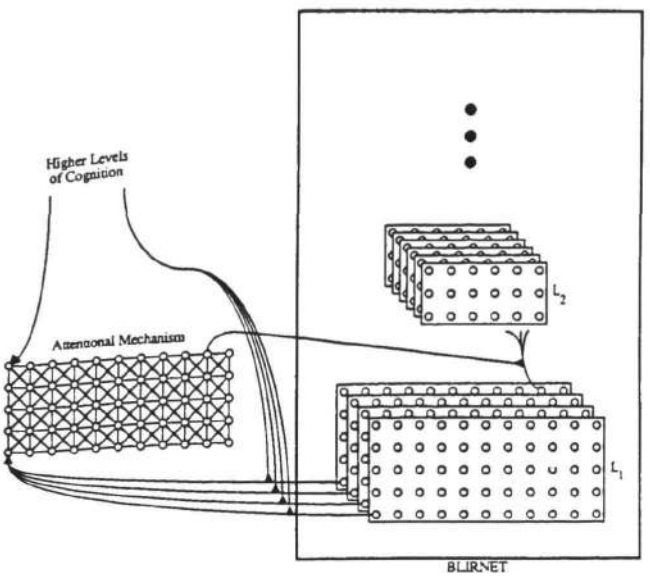

Figure 2. The attentional mechanism and its relationship to BLIRNET.

letters and bits of words, regardless of retinal location. The other processing modules of MORSEL have similar input-output properties: a retinotopic input of elementary features, and a location-independent output representation of high-level conjunctive features.

THE ATTENTIONAL MECHANISM

What might an attentional mechanism look like in the context of MORSEL? I propose a simple mechanism, one that directs a "spotlight" to a particular region of the retina (e.g., Crick, 1984; Eriksen & Hoffman, 1973; Posner, 1980; Treisman & Gelade, 1980). The attentional spotlight serves to enhance the activation of low-level retinotopic features within its bounds relative to those outside. As activity is propagated through BLIRNET and the other modules, the highlighted region maintains its enhanced status, so that in the output layer of the module, units appropriate for the attended item(s) tend to become most active as well. Consequently, these units will dominate the pull-out net competition, causing the attended item(s) to be selected. In this way, the AM allows preferential processing of attended stimuli.

The Attentional Mechanism as a Filter

The AM (Figure 2) is a set of units arranged in a retinotopic map in one-to-one correspondence with the input layer (denoted L_1) of BLIRNET. Activity in an AM unit indicates that attention is focused on the corresponding retinal location and serves to *gate the flow of activity* from L_1 to the second layer (denoted L_2) of BLIRNET. Specifically, the activity level of an L_1 unit in location (x,y) is transmitted to L_2 with probability $\xi+(1-\xi)a_{xy}$ (the *transmission probability*), where a_{xy} is the activity level of the AM unit in location (x,y) and has range [0,1], and ξ is a scaling parameter with a value of approximately .25. As long as ξ is greater than zero, the AM serves only to *bias* processing; it does not absolutely inhibit activations from unattended regions (similar to the Norman and Shallice, 1985, model).

As one might expect, highly familiar stimuli outside the focus of attention can work their way through the system better than other stimuli. To illustrate this point, BLIRNET was tested midway through training on an isolated letter recognition task. Some letters were recognized better than others: X was detected in every location and in the context of virtually any other simultaneously-presented letters, H was less consistently detected, and F even less so. Taking stability of detection to be an indication of familiarity, one might predict that performance on X should suffer less than performance on H, and H less than F, when attention is removed. This prediction is confirmed by Figure 3a. Performance here is measured as the ratio of the activation level of the target letter to the activation level of the maximally active nontarget letter, averaged over thirty presentations of the target. When this ratio falls below 1.0, the target cannot be discriminated from the nontargets. X is discriminable as long as the transmission probability is greater than .1, H .3 and F .8. Thus, BLIRNET is able to recognize familiar stimuli based on fewer perceptual features than less familiar stimuli. In other words, focal attention is less critical for highly familiar stimuli.

To further illustrate the filtering properties of the AM, BLIRNET was tested on L and G presented simultaneously. Attention was varied from being fully divided (i.e., the transmission probability was 1.0 for both letters) to being focused solely on the L (i.e., the transmission probability was 1.0 for L and 0.0 for G). Figure 3b shows that by concentrating attention on L, its relatively weak response can be improved dramatically, although this improvement is matched by a corresponding decrement in the response to G. Thus, inter-item crosstalk is reduced by focusing attention on one item. (In this example, the target:spurious activity ratio is not an absolute measure of discriminability. Because there are two stimuli, what matters for recognition are the *two* most active units. Even if a target has a ratio less than one, it may still be the second most active unit.)

System Dynamics

In the previous section, I described the manner in which a given AM state influences processing in MORSEL. In this section, I turn to the issue of how this state is computed. I begin by assuming external sources of knowledge are available that offer suggestions about where to focus. Sometimes these suggestions will conflict with one another; the task of the AM is to resolve such conflicts and construct an attentional spotlight centered on the selected location.

The AM units are interconnected to form a relaxation network that settles into states having a single, convex region of activation. The activity of each unit is updated over time as follows:

$$a_{xy}(t+1) = f\left[\mu \sum_{i=-1}^{1} \sum_{j=-1}^{1} a_{x+i\ y+j}(t) - \theta \sum_{\substack{i \neq x \\ j \neq y}} a_{ij}(t) + ext_{xy}(t)\right],$$

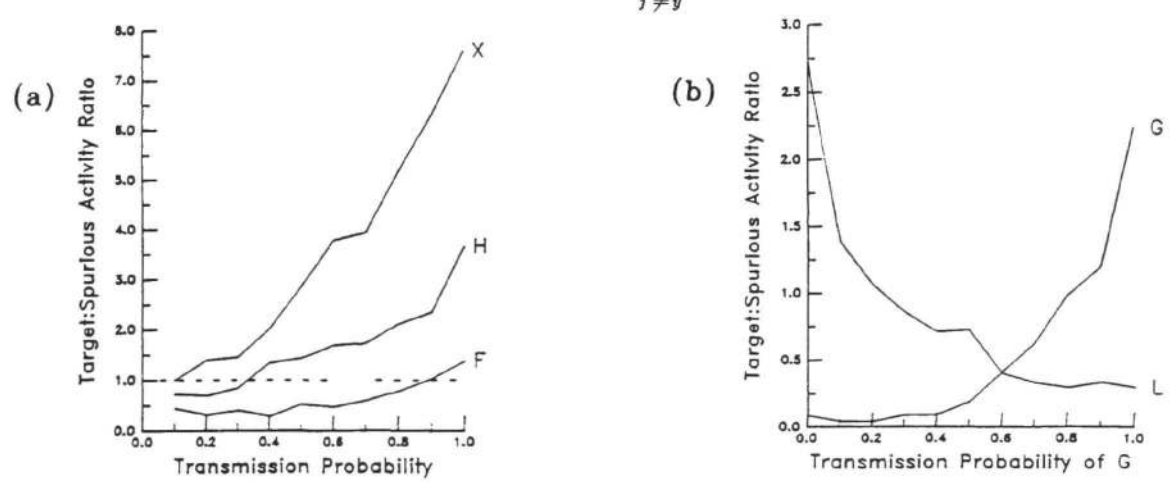

Figure 3. (a) Mean ratio of target activation to the maximum spurious (nontarget) activation for X, H, and F presented in location (14,2) as a function of transmission probability, averaged over thirty presentations. (b) Mean ratio of target activation to the maximum spurious activation for L in location (11,2) and G in location (20,2) as a function of attention to the G, averaged over thirty presentations.

where $a_{xy}(t)$ is the activity of AM unit in location (x,y) at time t, μ and θ are adjustable constants, $ext_{xy}(t)$ is an external input to location (x,y) at time t, and $f[x]$ is an identity function with saturation points at zero and one. The first term in the activation function encourages contiguous regions of activity by pushing each unit to take on the average value of itself and its eight spatial neighbors. (A neighbor is assumed to have activity level zero if it is outside the retinotopic array.) If μ is 1/9, an exact average is computed; as a result, activation levels fade with increased distance from the center of activity. If μ is larger, however, the boundary between active and inactive regions are sharpened, so that a unit will tend to be fully on if its neighbors are on or off otherwise. The second term in the activation function limits the total activity in the network by causing each unit to inhibit all others, with θ controlling the degree of inhibition. If several discontinuous regions are simultaneously active, this term serves to suppress all but the most active region. The third term allows external sources of knowledge to drive activity in the network.

Guiding the Spotlight

These external knowledge sources can be dichotomized into two classes: *data driven* and *conceptually driven*. To consider a simple case of a "data driven" source, attention should be drawn to objects but not empty regions in the visual field. This property is incorporated into the AM by having each L_1 unit project to its corresponding AM unit (Figure 2). Similar connections to the AM should be made from the elementary feature maps of other modules, e.g., maps detecting color, texture boundaries, and motion. Through these connections, attention can be captured by such varied stimuli as an intense or flashing light, object motion, or an odd element against a homogeneous background. Further control is required, however: the mere presence of any feature should not cause an attentional shift willy nilly; attention is dependent on higher-level expectations and task demands. For example, in the task of detecting a "–" in a display of oriented line segments, one would like for only the "–" features to trigger attention. I thus propose that higher levels of cognition (*HLC*) can modulate the effect of each feature type on the AM, allowing only the features of interest to capture attention. Mechanistically, this is not difficult to implement: HLC simply need to gate the connections from each feature type in L_1 (and other such feature maps) to the AM.

Besides data-driven guidance, "conceptually-driven" guidance — direct control by HLC — is required in many situations, from reading, where text must be scanned from left to right, to a variety of experimental tasks where selection is based on location (e.g., a precue indicating the location of an upcoming target item).

If items of interest in the visual field vary in size, so must the spotlight. Empirical evidence confirms this intuition (Eriksen & Yeh, 1985; Laberge, 1983). Thus, it seems critical that HLC be able to influence not only the locus of the spotlight but also its diameter. The spotlight diameter is modulated by the parameter θ. Consequently, I assume that θ is dynamically regulated by HLC as a function of time and task.

SIMULATION RESULTS

I have implemented a simulation of the AM in which the human operator is allowed to specify the external inputs. Figure 4a presents a simple example in which two external inputs have been given, one at location (7,4) with value .2 and the other at (16,3) with value .3. Initially, activity levels of all AM units are reset to zero. Over time, the external inputs are copied into the activity of the corresponding AM units. Spotlights then begin to form around each stimulated location, but gradually activity in the region of (7,4) is suppressed, due to the fact that only one spotlight can be supported and the external input to (7,4) is smaller. By iteration 15, the network reaches equilibrium. Figure 4b shows another example with the same external inputs but θ decreased from .02 to .01. The resulting spotlight is about twice as large as in Figure 4a. One might be tempted to conclude that θ directly regulates the diameter of the spotlight, but the story is more complex, as the next example demonstrates.

In Figure 4c, the external inputs specify two blob-like regions, not individual points of activation as in the previous examples. This input pattern was constructed by presenting the stimulus WIX MUJ to BLIRNET (see Figure 4d), and counting the number of feature detectors active in each location of L_1. This sort of an input pattern might arise naturally on the AM if each L_1 unit fed activity into its

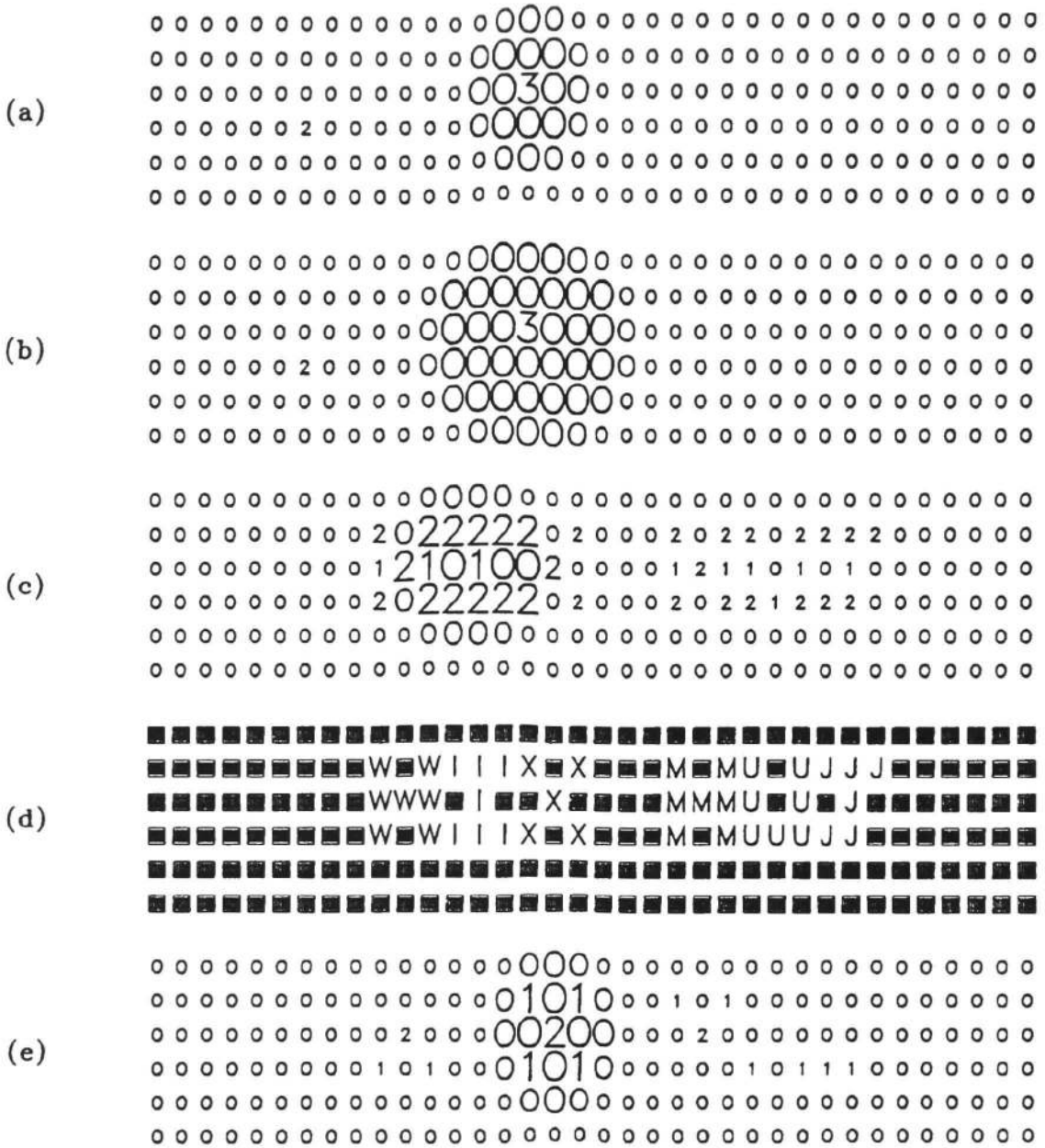

Figure 4. (a) Activity pattern in the AM at equilibrium resulting from two external inputs (μ=.22, θ=.02). The activity level of each AM unit is represented by the *size* of the digit in the corresponding position. The digit itself represents the magnitude of the external input (actually, ten times the input). (b) Activations in the AM resulting from two external inputs (μ=.22, θ=.01). (c) Activations in the AM resulting from external inputs concentrated in two regions, with slightly more input to the left region (μ=.22, θ=.02). (d) Location of the stimulus WIX MUJ that serves as input for Figures 4c and 4e. (e) Activations in the AM resulting from external inputs based on the \ and / features of the stimulus WIX MUJ (μ=.22, θ=.02). The location of the X is selected.

corresponding AM unit (as discussed earlier). The initial AM activity then reflects all bottom-up sources of information: attention is broadly tuned to include all items in the visual field. Over time, however, attention "narrows" on the left region — the site of WIX. This region is selected because its net external input is greater — 3.6 units of activity versus 3.5.

Although the same value of θ was used in Figures 4a and 4c, the spotlight in Figure 4c is larger. It appears that θ does not directly control the spotlight diameter. Roughly, θ can be thought of as a measure of the maximum distance allowed between two points of external activity in order for them to be enclosed within the spotlight: the larger θ is, the smaller the distance. This is a nice property of the system in that the spotlight should be on one object at a time, but it is unclear how to define the boundary of an object; with WIX MUJ, is the entire stimulus an object, is just the WIX, the X, or perhaps only one stroke of the X? θ provides one dimension along which an object's boundary can be characterized, namely, the maximum spacing between its components.

A final example of the operation of the AM is presented in Figure 4e. I have simulated the situation in which WIX MUJ is presented to BLIRNET and the L_1-AM connections are gated so that only the "\" and "/" feature maps trigger the AM. As a result, the letter X is selected. In this manner, higher levels of cognition can control which item will be selected, but only if the item has distinctive elementary features: pairs like W and M cannot be differentiated on the basis of elementary features.

THE ROLE OF ATTENTION

The AM serves MORSEL in four respects, suggesting the following roles of attention in visual information processing.

(1) *Controlling order of read out.* The AM allows MORSEL to selectively access information in the visual field by location.
(2) *Reducing crosstalk.* When items are analyzed simultaneously by MORSEL, interactions within the processing modules cause interference among items. By focusing attention on one item at a time, crosstalk can be reduced.
(3) *Recovering location information.* Remember that the output of BLIRNET — the letter-cluster representation — encodes the identity of a letter or word but not its retinal location; the operation of BLIRNET and the other modules factor out location information. However, because the current focus of attention reflects the spatial source of letter-cluster activations, the AM can convey the lost location information.
(4) *Coordinating processing performed by independent subsystems.* Each processing module operates independently of the others. Consequently, it is imperative to ensure that the results from the various modules are grouped appropriately. The AM allows this by guiding processing resources of all modules to the same spatial region. This function of attention seems analogous to that suggested by feature-integration theory (Treisman & Gelade, 1980).

EARLY VERSUS LATE SELECTION: WHERE DOES THE AM FIT IN?

A central issue in perceptual psychology over the past three decades has been the level at which attentional selection operates. Theories of attention can be dichotomized into two opposing views: *early* and *late selection.* Early-selection theories (Broadbent, 1958; Treisman, 1969) derive their name from the assertion that selection occurs early in the sequence of processing stages, prior to stimulus identification. In contrast, late-selection theories (e.g., Deutsch & Deutsch, 1963; Norman, 1968; Shiffrin & Schneider, 1977) posit that selection occurs late in processing, following stimulus identification. Additional properties go hand in hand with the central assumption of each theory (Pashler & Badgio, 1987). Early selection generally implies that (1) selection is based on low-level features such as stimulus location or color, (2) the processing system is of quite limited capacity, and (3) stimulus identification is necessarily serial. In contrast, late selection generally implies that (1) selection is based on high-level features such as stimulus identity, (2) the processing system is without capacity limitations, and (3) stimulus identification proceeds in parallel.

The view of attention presented by MORSEL is neither strictly early nor late selection. It agrees with late-selection theories in suggesting that multiple display items can be processed in parallel to a high level of representation, even to the point of making simultaneous contact with semantic knowledge (which occurs in the pull-out net). Further, selection via the pull-out net can be based on high-level — semantic or orthographic — features; this is accomplished by priming semantic units or letter-cluster units in the pull-out net to bias the pull out process. In other respects, however, MORSEL embodies an early-selection theory. First, the AM is an early selection device. It operates on a low-level representation, much in the spirit of the filtering and attenuation operations proposed by early-selection theories.

Second, the processing capacity of MORSEL is limited. If multiple items are analyzed simultaneously, interactions among the items can lead to damaging crosstalk; and there is the further problem that information about the location of each item is lost.

MORSEL thus shows characteristics of both early and late selection theories. Pashler and Badgio (1985, 1987) have proposed a similar hybrid view of attentional selection based on a large body of empirical work. Their view seems entirely compatible with MORSEL and the AM. I find it both surprising and exciting that MORSEL is in such close accord with the conclusions of Pashler and Badgio. MORSEL was not designed specifically to address attentional issues, yet it makes strong predictions concerning the nature of attentional selection. Furthermore, the hybrid view of attentional selection presented here seems like a possible resolution to the longstanding debate between proponents of early and of late selection.

In closing, I should note that Koch and Ullman (1985) have developed a related neurally-inspired model of the attentional spotlight. Their model is similar to the AM in that it consists of a topographic map in which units are activated to indicate the allocation of attention. Additionally, it operates by gating the flow of activity from a low-level input representation composed of elementary features. In Koch and Ullman's model, however, selection is performed by a simple winner-take-all network. This results in a single point of activity, as compared to the distributed activity pattern produced by the AM. Their model is thus unable to adjust the diameter of the attentional spotlight. A further drawback of the Koch and Ullman model is that it is embedded in a serial processing system, capable of processing only one item at a time. Without a system like BLIRNET, their model is merely an early selection device. This brings up the point that it is not the attentional mechanism itself that determines whether the system as a whole is best characterized in terms of early or late selection, but rather how the attentional mechanism is integrated into the rest of the system. This is where MORSEL makes a distinct contribution to theories of attention.

REFERENCES

Broadbent, D. E. (1958). *Perception and communication.* London: Pergamon.

Crick, F. (1984). The function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences, 81,* 4586-4590.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70,* 80-90.

Eriksen, C. W. and Hoffman, J. E. (1973). The extent of processing of noise elements during selective coding from visual displays. *Perception and Psychophysics, 14,* 155-160.

Eriksen, C. W. and Yeh, Y.-Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 583-597.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4,* 219-227.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 371-379.

Mozer, M. C. (1987). Early parallel processing in reading: A connectionist approach. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 83-104). Hillsdale, NJ: Erlbaum.

Mozer, M. C. (1987). *The perception of multiple objects: A parallel, distributed processing approach* (Unpublished Doctoral Dissertation). University of California, San Diego.

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75,* 522-536.

Norman, D. A. and Shallice, T. (1985). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Ed.), *Consciousness and self regulation: Advances in research, Vol. IV.* New York: Plenum Press.

Pashler, H. and Badgio, P. C. (1985). Visual attention and stimulus identification. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 105-121.

Pashler, H. and Badgio, P. C. (1987). Attentional issues in the identification of alphanumeric characters. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 63-82). Hillsdale, NJ: Erlbaum.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3-25.

Shiffrin, R. M. and Schneider, W (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127-190.

Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psych. Review, 76,* 282-299.